Une interface pour l’exploitation de corpus arborés par des non-informaticiens : la plate-forme ScienQuest du projet Scientext

Achille Falaise* — Agnès Tutin — Olivier Kraif****

achille.falaise@imag.fr, agnes.tutin@u-grenoble3.fr, olivier.kraif@u-grenoble3.fr

** GETALP-LIG, BP 53, 38041 Grenoble cedex 9*

*** LIDILEM, Université Stendhal, UFR SdL, BP 25, 38040 Grenoble cedex 9*

RÉSUMÉ. *La communauté du TAL développe de nombreux corpus, souvent librement disponibles, disposant d’annotations riches mais difficilement utilisables pour des chercheurs non informaticiens. Si la communauté du TAL souhaite ouvrir ses corpus annotés à un public plus large, elle doit impérativement concevoir et déployer des interfaces simples, ce qui n’est pas un problème trivial. Dans cet article, nous réfléchissons, dans le cadre du projet Scientext, aux critères ergonomiques et aux méthodes permettant d’élaborer un système de requêtes facile d’accès et soulignons les limites de la plupart des outils existants. Nous présentons la plate-forme ScienQuest, conçue pour effectuer sans connaissances techniques préalables des recherches sur les parties textuelles, les parties du discours et les fonctions syntaxiques. Conformément à nos attentes, une première évaluation montre une préférence marquée pour les modes de recherche les plus simples. Au-delà du projet Scientext, l’environnement ScienQuest, conçu comme un outil générique, devrait permettre d’intégrer rapidement de nouvelles ressources textuelles libres.*

ABSTRACT. *The NLP community has developed many corpora with rich annotations but these resources are not easily accessible to researchers with little computer expertise. If the NLP community is eager to make available annotated corpora to a wider audience of non-specialists, it is imperative to design and develop user-friendly interfaces, which is not a trivial problem. In this article, in the framework of the Scientext project, we examine several criteria and methods in order to develop such an interface and we highlight the drawbacks of existing systems. We then present the ScienQuest system, dedicated to several kinds of linguistic queries : textual parts, part of speech, syntactic functions. As expected, a first evaluation shows that simple and assisted query modes are preferred to complex query languages. Beyond the Scientext Project, the ScienQuest environment, developed as a generic tool, is planned to be used with various free textual resources.*

MOTS-CLÉS : *interface, annotations, corpus arborés, écrits scientifiques.*

KEYWORDS: *interface, annotation, treebanks, scientific writings.*

1. Introduction

Les corpus textuels actuellement disponibles sont fréquemment enrichis par différents types d'annotations linguistiques dont nous présentons ici les plus fréquentes :

– annotations morphosyntaxiques, presque partout disponibles pour la plupart des grands corpus de référence comme *Frantext*, le *British National Corpus*, le *Contemporary Corpus of Contemporary American English*, Ruscopora, etc. ;

– annotations syntaxiques, déjà moins fréquentes, comme par exemple le French Treebank (Abeillé *et al.*, 2003) ou l'Arboratoire (Bick, 2005 ; Salmon-Alt *et al.*, 2004) pour le français ;

– annotations discursives, par exemple l'annotation de la coréférence dans des corpus du français (Tutin *et al.* 2000 ; Salmon-Alt 2002), ou l'annotation des structures de discours (Péry-Woodley *et al.* 2009) ;

– annotations sémantiques, en particulier avec des désambiguïsations sémantiques et lexicales, en exploitant par exemple le réseau Wordnet[1].

Des annotations structurelles suivant les recommandations de la *Text Encoding Initiative* sont également souvent proposées. Ces informations sont utiles pour effectuer des études contrastives entre catégories de documents, ou entre les différentes parties des documents.

Cependant, les outils d'exploration de corpus annotés restent trop souvent complexes à utiliser, *a fortiori* pour des utilisateurs non initiés à la linguistique-informatique. L'ergonomie et la facilité d'utilisation des outils sont néanmoins des enjeux majeurs en TAL, surtout si l'on souhaite diffuser des traitements et des annotations linguistiques complexes dans la communauté des linguistes, en particulier pour outiller des ressources libres. Pour élargir le nombre d'utilisateurs des corpus annotés, nous pensons qu'il est essentiel de développer des outils d'exploration de corpus faciles à manipuler mais offrant des fonctionnalités riches, prenant en compte des annotations de haut niveau (syntaxe, partie textuelle, etc.) et permettant des requêtes élaborées permettant à l'utilisateur de croiser différents critères. C'est ce qui nous a amenés à proposer un environnement de recherche simple, adapté aux linguistes, didacticiens, lexicographes ou épistémologues[2] dans le cadre du projet ANR Scientext, qui intègre des corpus annotés aux plans structurel, morphosyntaxique et syntaxique. Cet outil, baptisé ScienQuest[3], a depuis été adapté pour accueillir d'autres corpus.

1. Cf. *Wordnet Gloss Desambiguation Project*, *Princeton University*: *Semantically annotated gloss corpus* (2008) : http://wordnet.princeton.edu/glosstag.shtml

2. *A priori*, les spécialistes des domaines scientifiques ne sont pas visés en priorité par l'application, dans la mesure où la taille des corpus recensée est encore modeste.

3. ScienQuest est utilisable en ligne sur http://scientext.msh-alpes.fr .

Après une réflexion sur les critères ergonomiques et les méthodes permettant d'élaborer une interface conviviale dans le cadre du projet Scientext, nous montrons les limites de quelques systèmes d'interrogation de corpus de référence étiquetés morphosyntaxiquement, et de corpus arborés, principalement pour le français. Nous détaillons ensuite les fonctionnalités de notre outil ScienQuest (recherche sémantique, libre, et avancée), et effectuons enfin un premier bilan de son utilisation dans le cadre du projet Scientext, après quelques mois d'existence publique qui révèle une nette préférence des utilisateurs pour les modes de recherche simples.

2. Utilisateurs, scénarios et critères d'utilisabilité de corpus annotés linguistiquement : application au projet Scientext

2.1. *Définition de profils d'utilisateurs*

Qui sont les utilisateurs des corpus annotés ? Cela peut concerner des publics variés : TAListes en train de construire des grammaires, linguistes étudiant la distribution d'un phénomène linguistique, littéraires en pleine étude de style, apprenants d'une langue étrangère souhaitant vérifier l'usage d'un terme ou d'une tournure de phrase... sont autant d'utilisateurs potentiels. Lors de l'élaboration d'un outil, il est primordial de savoir si l'outil envisagé est facilement utilisable par le public visé. Un critère essentiel est l'expertise informatique de l'utilisateur. En effet, les outils d'exploitation de corpus sont généralement créés par des informaticiens ou des TAListes, c'est-à-dire des utilisateurs ayant un bon niveau d'expertise informatique, qui n'ont pas nécessairement conscience des difficultés d'un utilisateur peu expert. En réalité, comme nous le verrons par la suite, peu d'outils sont véritablement destinés aux non-spécialistes, et nous essayons de répondre à ce besoin dans le projet Scientext.

2.2. *Le projet Scientext*

La réflexion sur les interfaces a été élaborée dans le cadre du projet Scientext, qui visait un double objectif ; d'une part, du point de vue linguistique, une étude des marques du positionnement de l'auteur (cf. plus bas), d'autre part, dans une optique plus ingénierique, la constitution d'un corpus représentatif d'écrits scientifiques annotés linguistiquement et d'un ensemble d'outils permettant d'interroger en ligne ce corpus. C'est dans ce dernier cadre qu'une réflexion sur l'interface d'utilisation du corpus a été menée.

Ce projet a réuni plusieurs laboratoires : a) le Laboratoire de Linguistique et de Didactique du Française Langue Étrangère et Maternelle, pilote du projet, b) le

laboratoire Littérature Langage Société, de l'Université de Chambéry, c) l'équipe Linguistique de Corpus, Université de Bretagne Sud.

La modélisation linguistique autour du thème du positionnement visait à proposer une articulation entre les niveaux rhétorique, sémantique (et énonciatif) et lexical. Dans l'étude du positionnement dans l'écrit scientifique, plusieurs entrées ont été identifiées. Tout d'abord, le thème du contexte scientifique, du cadre théorique et des références propres à un auteur ou à une équipe, a donné lieu à des travaux sur les marques de la filiation intellectuelle et du cadrage théorique (Grossmann *et al.*, 2009 ; Boch *et al.*, 2007). Par ailleurs, des travaux ont été élaborés autour du thème du parti pris, du jugement, de l'évaluation (Tutin, 2010b). Enfin, le thème des choix propres et du raisonnement a été abordé à travers les verbes de positionnement et le fonctionnement évidentiel des verbes de perception (Grossmann et Tutin, 2011). Dans notre approche, nous avons fait l'hypothèse d'un fonctionnement relativement stéréotypé de l'expression du positionnement – hypothèse en grande partie vérifiée par la suite – s'exprimant à travers une phraséologie repérable dans les corpus et facilement modélisable. Nous souhaitions pour ces études pouvoir extraire cette phraséologie en prenant en compte les disciplines, les différentes parties textuelles et comparer les genres des écrits scientifiques (par exemple, les différences entre articles de recherche et thèses) et avons développé la plate-forme ScienQuest dans cette optique. Les corpus constitués dans ce projet, les annotations linguistiques et les modes de recherche prévus sur ces données sont décrits dans la section 4.

2.3. *Besoins des utilisateurs peu experts en informatique*

Dans le cadre du projet Scientext, nous avons procédé en deux grandes phases pour élaborer notre outil de recherche :

1. une enquête de besoins auprès des utilisateurs visés ;
2. l'élaboration d'un prototype et son évaluation à plusieurs reprises.

L'étape 2 a été répétée régulièrement, le prototype étant adapté à chaque itération afin de corriger les problèmes constatés et d'introduire progressivement de nouvelles fonctionnalités.

2.3.1. *Enquête auprès d'utilisateurs potentiels*

Une première enquête a eu lieu en mars 2008 auprès d'une quinzaine de chercheurs et d'étudiants en linguistique, didactique et communication, issus de plusieurs laboratoires de recherche (LIDILEM, LLS, LiCoRN et GRESEC). Il s'agissait de réfléchir à l'élaboration d'une interface permettant d'interroger un

premier échantillon du corpus Scientext constitué d'écrits scientifiques analysés syntaxiquement et structurellement. Plusieurs utilisateurs ont clairement indiqué leur difficulté à utiliser les outils disponibles alors, y compris des outils assez classiques comme Frantext, pour lequel la recherche à l'aide d'expressions régulières est difficilement maîtrisée par des utilisateurs occasionnels. Nous leur avons demandé d'exprimer à l'aide d'exemples leurs besoins en termes de recherches dans les corpus du projet Scientext, puis de se prononcer sur une première maquette d'interface, et en particulier sur son accessibilité pour des utilisateurs peu experts en informatique.

En dialoguant avec ces utilisateurs, un scénario générique a ainsi pu être mis en place. Celui-ci proposait une recherche en trois étapes :

1. **définition d'un sous-corpus** à partir des textes du corpus ;
2. **construction d'une requête** dans un mode assisté ;
3. **affichage des résultats, tri et exportation**.

En termes de recherches, les utilisateurs ont surtout mis en avant des besoins assez simples, d'ordre lexical, avec des combinaisons de critères, par exemple trouver tous les adverbes en *-ment* dans les textes ; à ce stade de l'étude des besoins, les relations syntaxiques présentes dans le corpus n'ont pas été mentionnées par les utilisateurs.

En termes d'exploitation, c'est essentiellement un affichage sous forme de concordancier qui était demandé, Cette demande s'accompagnait de la possibilité d'effectuer des tris et d'exporter dans des formats courants (CSV, XLS, HTML) les résultats.

2.3.2. *Élaboration d'un prototype et évaluation*

À la suite de la première enquête, un premier prototype a été élaboré puis évalué. Cette étape a été itérée plusieurs fois[4].

Une première évaluation (et mise à disposition) a notamment eu lieu sur la version 0.8 de ScienQuest. Cette version comportait la sélection de sous-corpus, la recherche en mode avancé (par langage de requête), l'affichage KWIC[5] des résultats, et le calcul de statistiques ; les recherches sémantiques (prédéfinies) et libres (à l'aide d'un assistant) n'étaient pas encore disponibles (voir partie 4 pour la description détaillée des fonctionnalités de ScienQuest). Certains utilisateurs, ayant du mal à formuler des besoins en termes formels, ont demandé à travailler sur les collocations, en lien avec les thèmes linguistiques du projet. Ces premières évaluations ont montré l'intérêt de la sélection de sous-corpus par critères, et du

4. La première version stable de ScienQuest était la 0.9, il est maintenant en version 1.4.

5. KWIC : *KeyWord In Context*.

calcul de statistiques, mais le langage de requête était toujours jugé trop complexe par les utilisateurs, qui se cantonnaient généralement à des recherches très simples du type cooccurrence de deux lemmes contigus.

Une deuxième évaluation, intégrant un assistant pour un mode de recherche simple, a été proposée auprès de chercheurs internes au projet (LIDILEM, LLS) et externes (CECL[6] de Louvain). Ce mode de recherche libre et assisté a été accueilli très favorablement. En particulier, beaucoup d'utilisateurs ont alors commencé à travailler avec les relations syntaxiques à partir de cette version (notamment pour l'extraction de la phraséologie), alors que peu d'entre eux utilisaient les relations syntaxiques avec le langage de requête du mode avancé jugé trop complexe. Ainsi, à l'aide du mode simple, les utilisateurs ont pu mieux exploiter les possibilités offertes par les annotations du corpus. Des besoins supplémentaires sont apparus, comme le traitement de la syntaxe profonde (en particulier les passifs dont l'analyse avec Syntex n'est pas très intuitive) ou des recherches portant sur la ponctuation et non seulement les mots ; l'émergence de ces besoins témoigne d'une meilleure prise en main par les utilisateurs.

Ce cycle de développement « en spirale » de l'interface, en lien avec les retours des utilisateurs, a été renouvelé jusqu'à la version actuelle.

2.4. *Critères d'évaluation d'un environnement de recherche pour non-spécialistes*

À l'issue de ces consultations et expérimentations, nous avons pu mettre en évidence un ensemble de critères. Un environnement de recherche sur corpus convivial, facilement utilisable par des non-spécialistes, doit, selon nous, pouvoir répondre à plusieurs exigences, que nous nous sommes efforcés de prendre en compte dans l'élaboration de l'interface ScienQuest.

– **Absence de technicité.** Le système doit être utilisable sans connaissances préalables, en tout cas pour une première approche, d'un langage de requête spécifique ou d'un langage de balisage comme XML. Les éléments spécifiques ou techniques devront être transformés par des valeurs préétablies intégrées dans des ascenseurs ou des listes à cocher. Les termes employés devront être le moins techniques possible, ce qui constitue une véritable gageure pour des annotations linguistiques complexes comme les annotations syntaxiques ou les annotations discursives.

– **Rapidité et facilité d'emploi.** Le système doit être rapide et simple d'emploi. L'utilisateur ne doit pas avoir à parcourir de documentation, en tout cas pour une

6. *Centre for English Corpus Linguistics*, Université catholique de Louvain.

utilisation standard. L'usager sera guidé dans sa démarche tout au long du processus.

– **Expressivité et progressivité.** Le mode assisté doit permettre d'exploiter le mieux possible la richesse de l'annotation. Il est intéressant de prévoir une progressivité d'un mode simple à un mode plus complexe, dans une démarche didactique. Il est évidemment impossible de proposer en mode simple assisté toute la richesse qu'offre un langage de requête complexe. Il sera néanmoins intéressant d'amener l'utilisateur « en douceur » à cette progressivité.

De plus, il ne faut pas avoir un point de vue trop simpliste sur les interfaces réalisées. Une interface graphique sophistiquée, telle que TigerSearch, peut, par exemple, se révéler plus complexe pour le linguiste qu'une interface plus austère, avec un mode commande simple. Pour l'élaboration des interfaces, c'est le pragmatisme qui doit prévaloir, avec des évaluations et des retours réguliers auprès des utilisateurs finaux.

3. Évaluation des outils existants

Comme nous l'avons signalé plus haut, la plupart des outils de recherche sur corpus restent relativement complexes à utiliser, et peu d'entre eux permettent selon nous d'exploiter toute la richesse des annotations. Nous examinons ici les interfaces proposées pour quelques corpus de référence étiquetés et arborés.

3.1. *Interfaces pour les corpus étiquetés morphosyntaxiquement*

À l'heure actuelle, il existe de nombreux environnements de recherche pour les corpus avec un étiquetage morphosyntaxique, mais peu d'entre eux apparaissent vraiment facilement utilisables pour des linguistes non spécialistes du TAL, en tout cas pour les fonctionnalités linguistiques les plus avancées. Nous observons cela pour la plupart des interfaces utilisées pour les grands corpus de référence du français ou de l'anglais comme le *British National Corpus* (BNC) en ligne, le *Corpus of Contemporary American English* (COCA) ou *Frantext* pour ne citer que les plus connus (cf. tableau 1 ci-dessous).

La plupart de ces corpus utilisent des langages de requête complexes, plus ou moins normalisés, plutôt qu'un ensemble de valeurs préétablies[7]. L'interface en ligne du BNC recourt ainsi au Corpus Query Language (CQP), qui n'est pas excessivement compliqué, mais requiert toutefois une connaissance des expressions régulières. Une interface simplifiée a été proposée par M. Davies pour le même

7. Dans des cases à cocher, par exemple.

corpus ainsi que pour le COCA, d'une simplicité d'accès tout à fait remarquable. On notera toutefois qu'il n'est pas prévu dans l'interface graphique de proposer à la fois une contrainte sur le mot et/ou le lemme et la partie du discours. L'ergonomie de la base Frantext, le grand corpus littéraire de référence du français, a été considérablement améliorée dans les dernières années, mais la recherche sur les parties du discours et les lemmes s'effectue toujours à l'aide d'un langage de requête spécifique et complexe, peu facile à manipuler pour les utilisateurs occasionnels. Parmi les grands corpus de référence (dont nous ne pouvons faire l'inventaire ici), certains utilisent toutefois des interfaces à la fois puissantes et conviviales, comme Ruscorpora[8], le corpus national russe, construit par l'Académie russe des sciences. Ce grand corpus du russe de 145 millions de mots librement accessible en ligne est finement annoté morphosyntaxiquement, partiellement analysé syntaxiquement et désambiguïsé sémantiquement. Son interface permet à la fois d'effectuer des recherches assistées et des requêtes complexes à l'aide d'un langage de requête.

Corpus	**URL**	**Mode de requête pour les étiquettes morphosyntaxiques et les lemmes**	**Exemple en français et équivalent en anglais *ce(s) problème(s) être ADJ***
BNC	www.natcorp.ox.ac.uk/	Langage Corpus Query Language (utilisé dans le BNC et Word Sketch Engine) et de nombreuses autres applications Interface simplifiée sur le site de Mark Davies : http://corpus.byu.edu/bnc/	[lemma="this"] [lemma= "problems "] [lemma= "be"] [tag= "ADJ.*"]
COCA	corpus.byu.edu/coca/	Langage de requête « maison » proche de CQL. Interface simplifiée.	[this] [problem] [be] [j*]
Frantext	www.frantext.fr/	Interface simplifiée pour les recherches simples. Langage de requête propriétaire pour les lemmes et les parties du discours puissant mais particulièrement complexe.	&e(c=(ce|ces)) &e(c=&mproblème) &e(c=&cêtre) &e(g=A)

Tableau 1. *Comparatif des modes d'interrogation de quelques corpus de référence*

Enfin, pour la recherche sur un corpus étiqueté morphosyntaxiquement en français, citons une interface qui nous paraît particulièrement intéressante, celle du système Elicop (Mertens, 2002), qui permet d'interroger des corpus de français oral

8. http://www.ruscorpora.ru

transcrit (dont le corpus d'Orléans). Ce système, qui a été une source d'inspiration pour notre interface, est basé sur un formulaire facile à remplir ne nécessitant pas de connaître un langage de requête (cf. figure 1). Le système ne permet toutefois pas de restreindre le corpus d'étude et est en outre limité à une recherche sur quatre mots ; les relations syntaxiques ne sont pas non plus prises en compte.

Search	Word 1	Word 2	Word 3	Word 4
Word Cat	- conditionnel	Adverb	- participe	Any
Lemma	avoir			
Form				

Figure 1. *L'interface de requête du projet Elicop : recherche sur le verbe* avoir *au conditionnel suivi d'un adverbe et d'un participe*

3.2. *Interfaces pour les corpus arborés*

Peu de corpus arborés sont actuellement disponibles pour le français, et encore moins en ligne. Le Corpus Arboré de Paris 7 (French Treebank) (Abeillé *et al.*, 2003), un corpus d'un million de mots de textes journalistiques, annotés en constituants, est disponible pour des travaux de recherche mais non consultable en ligne. Parmi les corpus arborés du français consultables en ligne, il n'existe, à notre connaissance qu'un corpus, l'Arboratoire, développé par Bick (2005) et Salmon-Alt (2002) dans le cadre du projet VISL et qui propose une analyse syntaxique dans le cadre de la grammaire de contrainte (*constraint grammar*) pour douze langues européennes. Les corpus du français (non corrigés manuellement) sont interrogeables en ligne grâce à un environnement d'étude, qui s'appuie toutefois sur un langage de requête complexe, et apparaît donc difficilement utilisable par un non-spécialiste. Le langage de requête utilisé, Tgrep2[9], qui est aussi utilisé pour l'interrogation du PennTreeBank en ligne[10], n'est pas véritablement accessible aux non-spécialistes.

Tigersearch (Lezius et König, 2000) est l'un des seuls environnements graphiques (mais hors ligne) permettant d'interroger des corpus arborés (de type syntagmatique) mais l'outil, qui reste assez proche du langage de requête sous-jacent, n'est plus maintenu à l'heure actuelle.

---

9. Pour le corpus appelé l'Arboratoire : http://corp.hum.sdu.dk/tgrepeye_fr.html
10. http://www.ldc.upenn.edu/ldc/online/treebank/. La requête suivante :
VP << /^believe/ < (S < (/^NP/ !<< /[*]/ !< (-NONE- < T)) < (VP|AUX << to))
indique par exemple que l'on recherche les occurrences du verbe *believe* qui ont un complément infinitif avec un sujet non nul.

Parmi les outils interrogeant des corpus arborés, nous pouvons à nouveau citer la réalisation exemplaire du site Ruscorpora qui propose ici aussi une interface particulièrement facile d'usage. Le corpus syntaxique, qui repose sur une analyse de dépendance inspiré du modèle Sens-Texte (Mel'čuk, 1988), est assez facile à interroger. La requête s'effectue entre deux mots, reliés par une relation de dépendance, et le choix est entièrement guidé par des cases à cocher. Dans la figure 2, on effectue une requête sur les verbes qui ont le nom hypothèse (гипотеза) comme complément d'objet (premier complément). La formulation de la requête est extrêmement simple. Il faut cependant préciser que les modèles de dépendance, reliant par définition des mots plutôt que des structures, sont probablement beaucoup plus faciles à représenter simplement dans des interfaces, que les modèles à base de constituants, qui manipulent des catégories abstraites. Néanmoins, deux points peuvent poser problème pour un utilisateur non spécialiste. Premièrement, l'interface se présente comme un outil ouvert, constitué de formulaires indépendants, et ne propose pas clairement de scénario de recherche. Cela complique un peu la prise en main par un utilisateur novice (cf. notre critère de facilité d'emploi). Deuxièmement, les formulaires utilisés dans l'interface sont riches et affichent d'emblée toutes leurs fonctionnalités. Cette richesse peut effrayer un utilisateur s'estimant peu expert (cf. notre critère de progressivité).

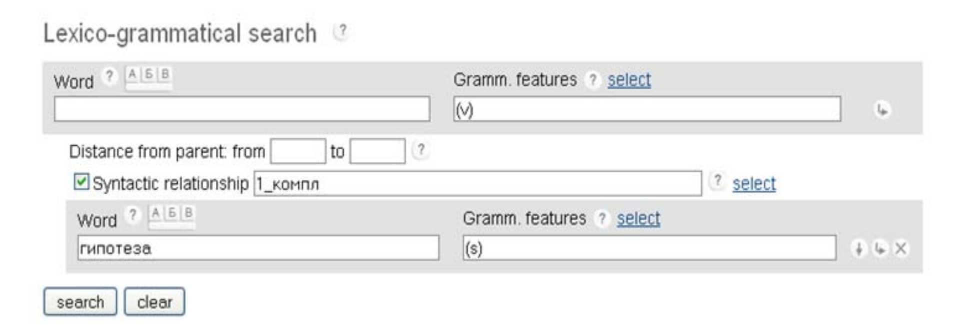


Figure 2. *Une requête syntaxique sur l'interface en ligne du corpus arboré Ruscopora : les verbes qui ont* гипотеза *(hypothèse) comme objet direct*

En conclusion, nous ne pouvons que déplorer le manque d'environnements en ligne conviviaux pour la recherche dans des corpus annotés en français, en particulier pour les corpus arborés. Lorsqu'ils sont simples à manipuler, les environnements sont assez limités dans les recherches. Inversement, les requêtes plus fines nécessitent souvent une vraie expertise de l'utilisateur concernant le langage de requête. En outre, ces systèmes sont dédiés chacun à un corpus en

particulier, et l'expérience acquise par les utilisateurs, parfois avec difficulté, n'est donc pas directement réutilisable sur d'autres corpus ; cela ne contribue évidemment pas à inciter les utilisateurs à faire l'effort d'apprendre à les maîtriser. Ces lacunes nous ont poussés à la réalisation de l'environnement ScienQuest, destiné à l'étude linguistique des écrits scientifiques.

4. Le système ScienQuest et son intégration dans le projet Scientext

ScienQuest est une plate-forme Web pour la consultation de corpus en ligne. Elle a été initialement développée pour les corpus du projet Scientext, mais est aussi actuellement utilisée pour d'autres corpus, en particulier dans le cadre du projet ANR Emolex[11]. La plate-forme est actuellement en version 1.4, version qui est présentée ici. ScienQuest est encore en développement, et pour l'instant son code et surtout sa documentation ne permettent pas une distribution *open source* efficace. Toutefois, une telle distribution est prévue sur le moyen terme.

4.1. *Architecture de la plate-forme*

La plate-forme ScienQuest (cf. figure 3) s'appuie sur le moteur de recherche ConcQuest, développé par Olivier Kraif (2008). Ce moteur est intégré dans un service Web de recherche, sur lequel s'appuie l'interface. ScienQuest est une application client-serveur, dont la partie serveur, codée en PHP, Perl et Prolog, fonctionne sur Unix/Apache, et la partie client, codée en HTML/Ajax est utilisable *via* un navigateur Web. L'interface de la plate-forme communique avec le service Web de recherche construit autour de ConcQuest par le biais d'une interface HTTP de type REST[12]. Cette interface est librement accessible et permet l'utilisation des corpus hébergés dans l'environnement ScienQuest par d'autres applications.

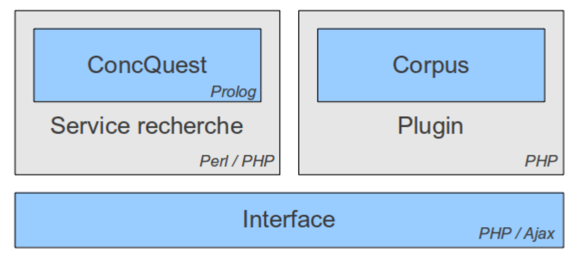


Figure 3. *Architecture de la plate-forme ScienQuest*

11. http://emolex.eu/

12. *REpresentational State Transfer*, une architecture basée sur les services Web, simple et largement utilisée.

4.1.1. *Généricité technique*

L'environnement d'exploitation développé dans le cadre du projet peut fonctionner avec plusieurs types de corpus, et a déjà été réutilisé en dehors du projet Scientext (notamment pour le projet ANR Emolex). L'ajout de corpus, toutefois, n'est pas une tâche triviale et doit être effectué par un spécialiste.

Les corpus sont intégrés à la plate-forme au sein d'extensions distinctes (*plugins*), ce qui permet d'intégrer des corpus de divers formats et des fonctionnalités spécifiques à un corpus donné. Actuellement, il est possible d'intégrer à la plate-forme des corpus écrits monolingues, répondant aux contraintes suivantes :

– soit non structurés, au format texte (UTF-8) ; soit structurés au format XML TEI Lite ;

– de langue et système d'écriture indifférent (testés avec les scripts latin et cyrillique) ;

– analysés, au choix, avec les analyseurs en dépendances comme Connexor, DeSR, Syntex ou XIP.

Le corpus doit être organisé en fichiers (normalement un texte par fichier) sauf s'il y a des parties sous-textuelles comme des introductions, conclusions, notes, titres, etc., qui doivent être dans des fichiers spécifiques.

La plate-forme ScienQuest a initialement été développée pour le corpus Scientext, analysé avec Syntex. C'est pourquoi la première étape de l'intégration consiste, le cas échéant, à convertir l'analyse au format Syntex. Il s'agit seulement d'une conversion superficielle, qui se résume souvent à renommer quelques balises XML ; les étiquettes (parties du discours, flexions, relations syntaxiques) ne sont pas converties ; elles sont détectées automatiquement au chargement du corpus. Actuellement, des scripts sont disponibles pour convertir les formats Connexor, DeSR et XIP vers Syntex. Cette conversion ne concerne que l'analyse, les en-têtes de fichiers sont extraits sans traitement spécifique et leur format est indifférent.

Pour pouvoir être intégré dans la plate-forme ScienQuest, le corpus doit ensuite être accompagné de trois éléments :

– un composant *plugin*, spécifique à chaque corpus, codé en PHP, capable de charger les en-têtes et de les convertir dans le format interne de ScienQuest. Un *plugin* « générique » est disponible pour les éléments de la TEI Lite utilisés dans le corpus Scientext, et peut être facilement adapté ;

– une description en XML de la structure du corpus : langue du corpus, langues disponibles dans l'interface, structure du corpus (types de textes, parties textuelles, etc.), catégories syntaxiques (parties du discours), etc. ;

– une liste de chaînes de caractères spécifiques au corpus : noms des catégories, des étiquettes, etc. L'interface supporte le multilinguisme, il est donc possible de fournir des listes de chaînes pour plusieurs langues.

Actuellement, outre les corpus publics du projet (cf. 4.2), cet outil est utilisé dans le cadre du projet ANR Emolex pour des corpus d'environ 200 millions de mots en allemand, espagnol, français (Connexor), anglais (XIP), et pour un corpus russe de 500 000 mots (DeSR).

Il est prévu de poursuivre l'intégration à ScienQuest de corpus libres. Des expérimentations ont commencé concernant des corpus libres annotés morphologiquement mais dépourvus d'annotations syntaxiques (utilisation du corpus de l'*Est Républicain* codé par le CNRTL, Bertrand Gaiffe et Kamel Nebhi ; création de corpus structurés en plusieurs langues issus de Wikipédia).

4.1.2. *Généricité fonctionnelle*

En ce qui concerne les fonctionnalités, l'environnement d'exploitation reste lié à son utilisation première dans le cadre du projet Scientext. La catégorisation des textes, les parties du discours, les relations syntaxiques, etc. s'adaptent à tout nouveau corpus, mais le scénario d'utilisation et les fonctionnalités restent ceux du projet d'origine. Ces fonctionnalités rencontrent cependant certaines limites, comme nous le verrons dans la dernière section.

4.1.3. *Gestion des droits des textes*

Un corpus peut être libre ou non. En fonction de la licence des textes, il peut être utilisable publiquement, ou bien être protégé par un login-mot de passe. En outre, dans ScienQuest, il est possible de combiner des textes libres et des textes non libres au sein d'un corpus : un visiteur non autorisé ne verra que les textes libres, alors qu'un utilisateur autorisé aura accès à tous les textes.

4.2. *Les corpus du projet Scientext : constitution et annotation*

Dans le cadre du projet Scientext, un corpus a été constitué et annoté[13]. Il comporte un ensemble de sous-corpus disponibles en ligne et interrogeables à l'aide des fonctionnalités décrites plus bas :

– un **corpus français d'écrits scientifiques** variés, comprenant 4,8 millions de mots dans 8 disciplines des sciences humaines, sciences expérimentales et sciences

13. Les corpus sont librement consultables sur http://scientext.msh-alpes.fr .

pour l'ingénieur, pour plusieurs genres d'écrits scientifiques (articles de recherche[14], communications publiées, thèses, mémoires d'HDR). Une partie de ce corpus est librement disponible pour la communauté scientifique[15] ;

– un **corpus anglais d'écrits scientifiques**, tiré du corpus BioMedCentral, principalement en biologie et en médecine, qui avoisine 13 millions de mots, qui a fait l'objet d'études lexicologiques (Williams et Millon, 2009) ;

– un **corpus anglais d'apprenants**, comprenant des travaux longs d'étudiants en anglais langue étrangère (1,1 million de mots) ;

– un **corpus expérimental de commentaires évaluatifs** pour une conférence de doctorants en sciences du langage (CEDIL) (cf. Boch *et al.*, 2011).

La question des droits pour l'accès aux corpus a été réglée diversement selon les corpus :

– pour les corpus d'apprenants et de commentaires évaluatifs, une autorisation a été demandée aux auteurs et les textes ont été anonymisés ;

– le corpus anglais BioMedCentral étant libre de droit, il a pu directement être utilisé dans notre projet ;

– pour le corpus français, deux types de conventions[16] ont été signées avec les auteurs et/ou les éditeurs des textes : la possibilité d'interroger le texte en ligne dans une limite de 200 mots (convention restreinte) ou la possibilité d'interroger le texte en ligne et de diffuser le document à l'aide d'une convention *Creative Commons* (respect du droit à la paternité ; pas d'utilisation commerciale ; pas de modification ; partage des conditions à l'identique).

Une large partie du corpus français (219 textes et 4,8 millions de mots, soit 75 % du corpus), annoté structurellement, est actuellement librement disponible pour la communauté des chercheurs[17].

Les corpus ont été annotés structurellement, avec une indication des principales parties textuelles (introduction, conclusion, titres, résumé, annexes, figures…) en suivant les recommandations de la TEI P5. Une annotation syntaxique automatique, sans révision manuelle, a également été effectuée grâce à l'utilisation de l'analyseur de dépendance Syntex, développé par Didier Bourigault (par exemple, 2007). La figure 4 donne un exemple d'analyse pour la phrase « *L'hémicorps gauche est*

14. Il s'agit généralement de publications académiques, sauf en médecine et en biologie, où ce type de publication est plus systématiquement en anglais ; pour ces disciplines, il s'agit de textes de quasi-vulgarisation à destination des professionnels.
15. Un corpus interne plus important est disponible pour les membres du projet. Il ne peut être librement mis en ligne car les droits d'auteur n'ont pas été négociés pour ce sous-corpus.
16. Les conventions ont été élaborées à l'aide d'un avocat spécialiste de la propriété intellectuelle, M[e] Josquin Louvier, de Grenoble.
17. Écrire à : scientext@u-grenoble3.fr pour obtenir le corpus.

préféré systématiquement au droit ». Il s'agit d'une annotation en syntaxe de dépendance de surface ; notamment, dans la phrase à la voix passive de l'exemple, on constate que le verbe plein n'est pas la tête de l'arbre de dépendance mais l'auxiliaire, et qu'il n'y a pas de relation syntaxique directe entre le sujet et le verbe plein.

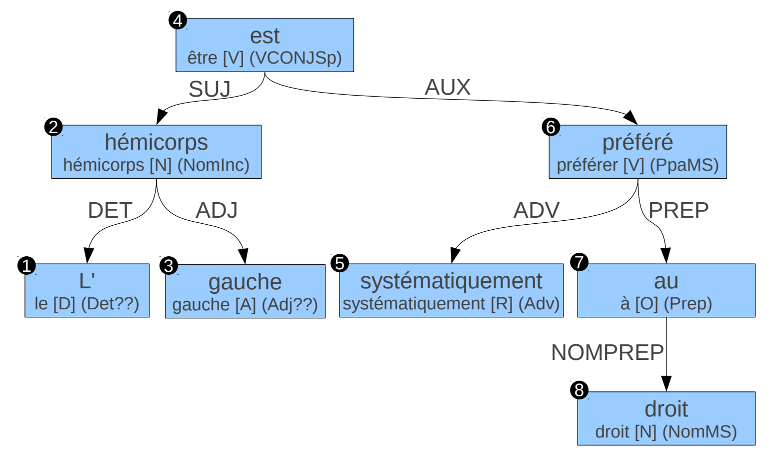


Figure 4. *Analyse syntaxique dans Syntex de l'énoncé « L'hémicorps gauche est préféré systématiquement au droit »*

Ces différents types d'annotations (structurelles, morpho-lexicales, syntaxiques) sont bien entendu exploités dans l'interface.

4.3. *Fonctionnalités de ScienQuest*

L'interface est bâtie selon le cheminement suggéré par notre échantillon d'utilisateurs. Nous avons fait le choix d'une approche segmentée en tâches simples et ordonnées, afin de guider l'utilisateur tout au long de la manipulation de l'outil. L'étape la plus complexe étant la construction d'une requête de recherche dans les textes, nous avons prévu pour cette étape plusieurs modes de recherche, en fonction du degré d'expertise de l'utilisateur : un mode sémantique (tous publics), un mode libre (nécessitant des notions de base en linguistique), et un mode avancé (pour les spécialistes du TALN). L'objectif est, pour les deux premiers modes, une interface utilisable dans un premier temps sans mode d'emploi, par un public non informaticien.

Nous avons fait le choix d'un cheminement en une succession d'étapes claires :

– choix d'un corpus ;

– travail sur ce corpus :

- définition d'un sous-corpus (voir section 4.3.1.),

- création d'une requête (voir section 4.3.2.),

- exploitation des résultats, à l'aide d'un concordancier et de statistiques (voir section 4.3.3.) ;

– exportation des résultats (aux formats CSV, XLS ou HTML).

Chacune des trois étapes principales est décomposée en sous-tâches plus simples, correspondant à une page de l'interface. Les exemples donnés ci-après sont fondés sur le corpus de textes scientifiques français du projet Scientext.

4.3.1. *Étape 1 : sélection d'un sous-corpus*

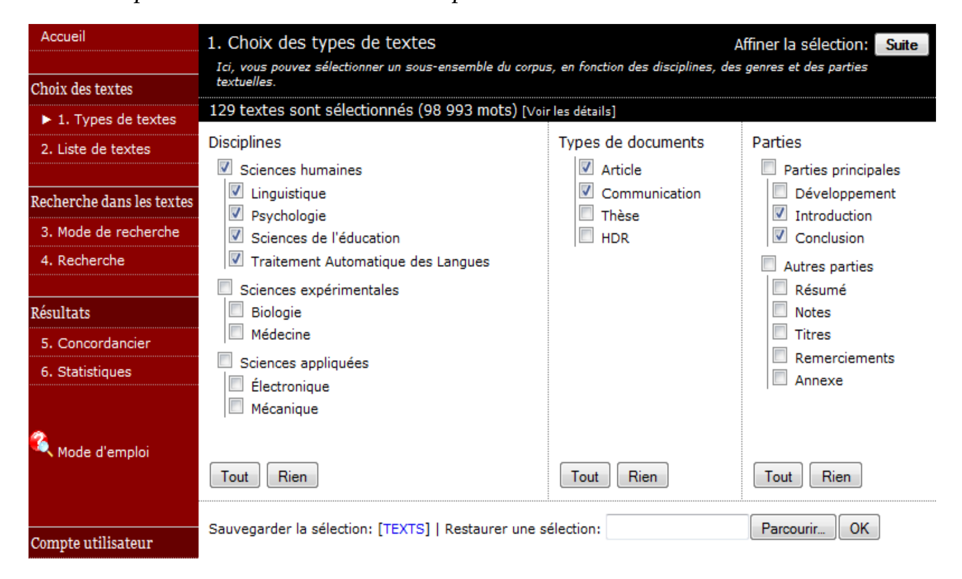


Figure 5. *Exemple de sélection d'un sous-corpus dans ScienQuest pour le corpus français de Scientext : les introductions et conclusions des articles de recherche et des communications en sciences humaines*

La première étape consiste soit à simplement accepter de travailler sur la totalité du corpus, soit à sélectionner un sous-corpus. Il est possible de combiner des groupes de textes préétablis suivant différents critères présentés dans des cases à

cocher : la ou les disciplines scientifiques, le type d'écrit scientifique, et la ou les parties textuelles. La figure 5 présente une sélection de communications et d'articles en sciences humaines où seules les introductions et conclusions ont été sélectionnées, soit presque 100 000 mots. Une fois le corpus sélectionné, l'utilisateur peut ensuite affiner la sélection en excluant certains textes. En outre, l'utilisateur peut sauvegarder sa sélection de textes dans un fichier local, qu'il pourra recharger dans une nouvelle session d'interrogation.

4.3.2. *Étape 2 : la recherche dans les textes*

Une fois le corpus délimité, l'utilisateur est invité à choisir entre trois modes de recherche : une recherche sémantique, une recherche libre guidée, et une recherche avancée. Quel que soit le mode choisi, le système produit toujours une requête complexe, dont le langage est présenté plus loin dans cette section (cf. recherche avancée). Cette requête complexe est compilée vers le langage de requête utilisé par le moteur de recherche ConcQuest, développé par (Kraif, 2008), qui effectue la recherche dans le corpus.

4.3.2.1. La recherche sémantique : recherche à travers des grammaires locales

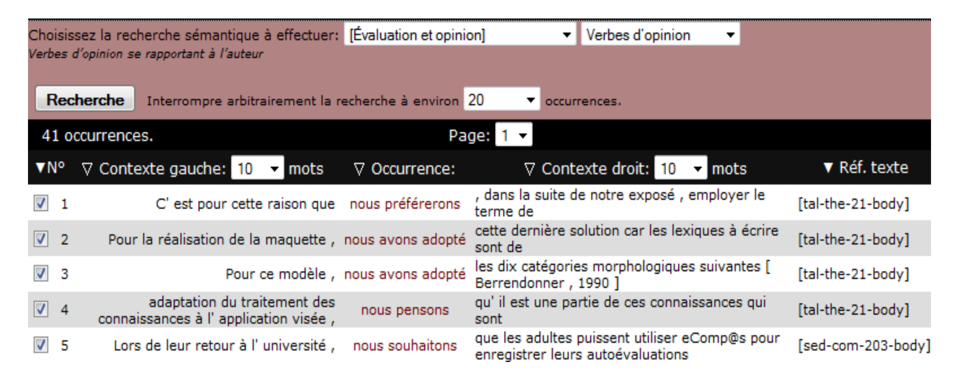


Figure 6. *Quelques résultats de la grammaire locale des verbes d'opinion*

Un mode d'interrogation innovant a été proposé dans l'interface pour les corpus de Scientext : une recherche d'expressions stéréotypées renvoyant à des fonctions sémantiques spécifiques, que nous appelons, à la suite d'autres linguistes comme Maurice Gross, des « grammaires locales ». L'élaboration de ces grammaires locales repose sur l'hypothèse d'un fonctionnement phraséologique de la langue, où l'expression de certaines fonctions sémantiques apparaît de façon privilégiée à travers certaines routines récurrentes (Tutin, 2010a). Ces grammaires locales, construites par les concepteurs du site, sont actuellement au nombre d'une quinzaine

et portent principalement sur le thème linguistique du positionnement de l'auteur, par exemple autour des verbes d'opinion ou des adjectifs d'évaluation. La figure 6 montre quelques résultats de la grammaire des verbes d'opinion. Il est bien entendu possible d'étendre ces grammaires locales à bien d'autres thèmes comme par exemple, les repérages des entités nommées, des expressions calendaires, etc. Dans le cadre d'applications didactiques, nous envisageons d'étendre ces grammaires à d'autres fonctions sémantiques et discursives, comme l'expression de la cause, le positionnement par rapport à d'autres auteurs, la formulation des problématiques. Ce type de recherche onomasiologique vise de nombreux types d'utilisateurs : non seulement des linguistes, mais également des apprenants en langue étrangère, des spécialistes des sciences de l'information, des épistémologues.

4.3.2.2. La recherche libre : utilisation d'un formulaire guidé pour les recherches morphosyntaxiques et syntaxiques

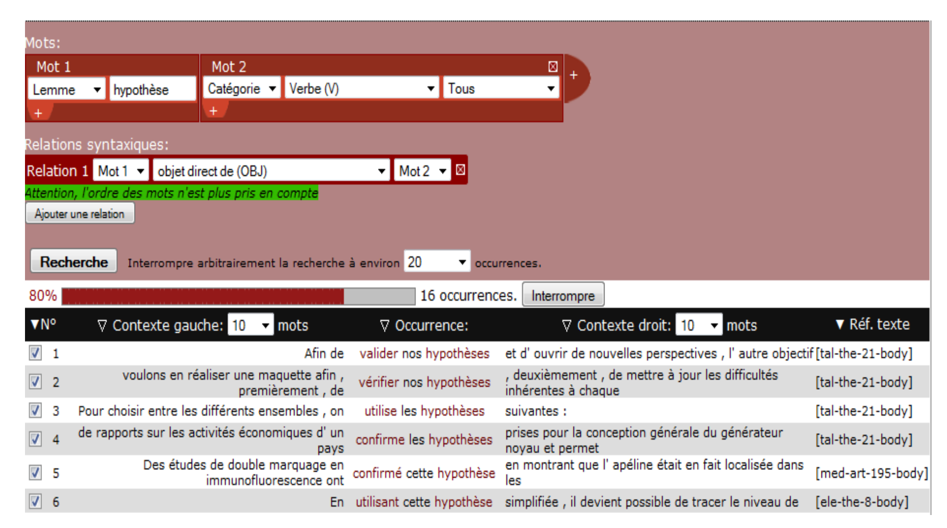


Figure 7. *Recherche sur les verbes ayant le lemme hypothèse comme objet direct*

Conformément aux principes développés en 2.4, ce mode de recherche répond aux critères d'absence de technicité, de rapidité de prise en main et de progressivité. Dans ce mode guidé, l'interface se présente d'abord de manière minimaliste, avec un champ de saisie pour une seule contrainte sur un seul mot. Des boutons permettent d'ajouter des mots et des contraintes sur les formes, les lemmes, les parties du discours (et éventuellement des sous-catégories). Les expressions régulières sont acceptées. En l'absence de relations syntaxiques, l'ordre des mots dans le formulaire est pris en compte lors de la recherche. Lorsque au moins deux

mots sont présents, la possibilité est offerte de spécifier une relation syntaxique entre ces mots. Si une relation est choisie, l'ordre des mots n'est alors plus pris en compte. Lors de la recherche, le contenu du formulaire est automatiquement converti en requête complexe. Par exemple, la figure 7 indique la recherche entre un verbe et le nom *hypothèse* lorsqu'il est objet direct de ce verbe.

Le mode libre permet d'effectuer des recherches suffisamment complexes pour la plupart des utilisateurs. Il est volontairement limité, afin de ne présenter qu'un sous-ensemble facilement compréhensible de fonctionnalités utilisables de façon intuitive, sans recourir à une documentation. Pour exploiter toute l'expressivité de l'outil de recherche, il faut passer en mode recherche avancée. Conformément à notre critère de progressivité, un bouton permet de passer de la recherche libre à la recherche avancée.

4.3.2.3. La recherche avancée : utilisation d'un langage de requête pour les corpus arborés

Le mode recherche avancée permet de créer directement une requête complexe, en suivant la documentation fournie. Ce mode est évidemment destiné aux utilisateurs spécialistes, linguistes familiarisés avec le TAL ou les traitements formels, informaticiens ayant des connaissances linguistiques. Les langages classiques d'interrogation de corpus, comme CQP, ne prenant pas en charge les relations de dépendances syntaxiques, il nous a fallu développer un langage spécifique fondé sur le langage du moteur de recherche ConcQuest. Ce langage de requête permet de spécifier des contraintes sur les mots (formes, lemmes, parties du discours, flexions), un ordre entre les mots, et des relations syntaxiques entre mots. Il est aussi possible d'utiliser des listes de mots et des variables.

Certaines fonctionnalités sont spécifiques au traitement des corpus arborés, en particulier la possibilité d'étendre les relations syntaxiques présentes dans le corpus. Par exemple, l'analyseur Syntex effectue une analyse syntaxique de surface et ainsi ne crée pas de relation de dépendance directe entre un verbe plein à un temps composé et son sujet, mais crée à la place une relation SUJ (sujet) entre le sujet et l'auxiliaire, et une relation AUX (auxiliaire) entre l'auxiliaire et le verbe. Dans ScienQuest, il est possible de définir une relation « sujet profond » qui prend en compte ce cas de figure.

Le tableau suivant synthétise les différents modes de recherche en fonction des types d'utilisateurs visés.

Mode de recherche	Exemples	Utilisateurs visés
Sémantique à l'aide de grammaires locales. Accès onomasiologique à partir de grammaires préétablies.	Verbes d'opinion. Formulation d'une hypothèse.	Tous types d'utilisateurs, y compris des utilisateurs n'ayant pas de connaissances linguistiques : apprenants en langue étrangère, spécialistes des sciences de l'information, épistémologues.
Libre et guidé Formulation guidée (lemmes, catégories et sous-catégories syntaxiques, relations syntaxiques de dépendance).	Suite de catégories syntaxiques. Verbes ayant *hypothèse* comme objet direct.	Linguistes ou utilisateurs maîtrisant les catégories et les fonctions syntaxiques.
Avancé Langage de requête utilisant des expressions régulières et les dépendances syntaxiques. Inégration de variables.	Création de grammaires locales ou de requêtes complexes avec des disjonctions de mots, des relations syntaxiques variées.	Linguistes familiarisés avec le TAL ou avec les traitements formels, informaticiens ayant des connaissances linguistiques.

Tableau 2. *Modes de recherche et utilisateurs visés*

4.3.3. *Visualisation des résultats, exportation des résultats et statistiques*

Les enquêtes auprès des utilisateurs ont montré que la visualisation des résultats et leur exportation étaient des éléments essentiels pour l'utilisation des corpus électroniques. Les linguistes, en particulier, souhaitent pouvoir éditer et retravailler les résultats dans des formats commodes.

Dans ScienQuest, les résultats des recherches sont affichés dans un format classique KWIC, paramétrable (voir la figure 7 pour un exemple). Ces résultats sont exportables en CSV, XLS et HTML. Comme avec Frantext, il est possible de « zoomer » sur des parties plus larges du texte, en respectant, lorsque l'information est disponible, le style du texte original (paragraphe, italique, etc.). Ce maintien du style d'origine apparaît tout à fait essentiel aux utilisateurs qui souhaitent prendre en compte l'information stylistique (structures de listes, italique…) qui n'apparaît pas dans le texte brut. En outre, les dépendances syntaxiques de la phrase affichée

peuvent être visualisées sous forme graphique, ce qui facilite la construction des requêtes complexes.

L'utilisateur peut en outre désactiver les résultats incorrects, qui ne seront pas exportés, ni comptabilisés dans les statistiques par la suite. Il s'agit d'une demande récurrente des utilisateurs, qui sont confrontés aux erreurs d'analyse du corpus, et veulent pouvoir filtrer les résultats incorrects.

De plus, des statistiques sur les occurrences trouvées sont disponibles, par exemple le nombre d'occurrences et le pourcentage des lemmes et des formes et leur distribution par discipline, genre textuel, partie textuelle, et par texte. Il s'agit d'une fonctionnalité encore peu présente dans les outils d'étude de corpus, particulièrement intéressante pour l'étude des structures rhétoriques dans l'écrit scientifique. Nous envisageons d'étendre ces fonctionnalités avec des statistiques textométriques et des mesures d'associations lexicales.

4.4. *Statistiques d'utilisation, évaluation et évolution du système*

Sur la période allant du début du lancement public du site en juillet 2010 à décembre 2011, 6 670 requêtes ont été effectuées (en 1 157 sessions) sur ScienQuest, concernant les corpus de Scientext. Le mode libre (guidé) est utilisé pour 73 % des requêtes, le mode sémantique (grammaires locales prédéfinies) pour 26 % et le mode avancé (langage de requête) pour 1 % ; cela démontre bien, selon nous, l'intérêt de ces deux premiers modes de recherche.

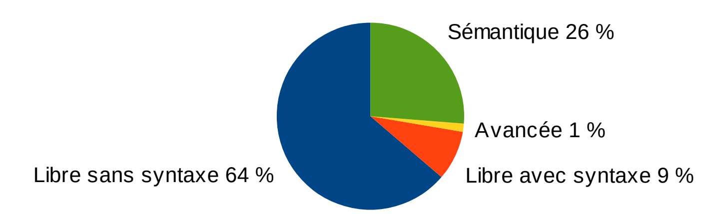


Figure 8. *Répartition des requêtes par mode. Pour le mode libre, on distingue les requêtes qui utilisent la syntaxe*

Les besoins exprimés dans l'enquête préliminaire (voir sections 2.3 et 2.4), étaient souvent très simples ; ils concernaient généralement des mots isolés, et faisaient rarement appel à des contraintes d'ordre syntaxique, qui présentent une complexité inhérente. Ces besoins se retrouvent dans les statistiques d'utilisation du mode libre. Ce dernier est privilégié par les utilisateurs, souvent pour des utilisations

simples. On relève ainsi que 30 % des requêtes libres ne contiennent qu'un seul mot, et 88 % n'utilisent pas les relations syntaxiques. Pourtant, si l'on regarde le détail des requêtes, on remarque que certaines d'entre elles gagneraient en souplesse grâce à l'utilisation des relations syntaxiques, alors que ces dernières sont pourtant négligées ; il s'agirait dont plus d'un problème de complexité que de besoin, qui semble réel. Par ailleurs, les informations flexionnelles, il est vrai assez pauvres dans le corpus Scientext, ne concernent qu'à peine 5 % des requêtes en mode libre ; les utilisateurs préfèrent souvent rechercher une ou plusieurs formes fléchies précises plutôt que d'utiliser ces contraintes. La notion de lemmes nous semble bien maîtrisée et appréciée par les utilisateurs, puisqu'ils sont utilisés dans 49 % des requêtes libres, alors qu'il ne s'agit pas du type de contrainte par défaut dans l'interface.

On observe, toujours sur la période allant de juillet 2010 à décembre 2011, une tendance à la progression du nombre d'utilisations de ScienQuest pour accéder aux corpus de Scientext (figure 9). Le nombre de requêtes par session croît sensiblement, ce qui tend à signaler une utilisation plus approfondie de l'outil au cours de sessions plus riches. Les pics de fréquentation peuvent s'expliquer par une utilisation en travaux dirigés dans le cadre universitaire, et par la publicité faite au cours de conférences (par exemple TALN en juillet 2011 ; colloque de l'AFLS en septembre 2011). *A contrario*, le nombre de visites décroît sensiblement pendant les périodes de vacances universitaires.

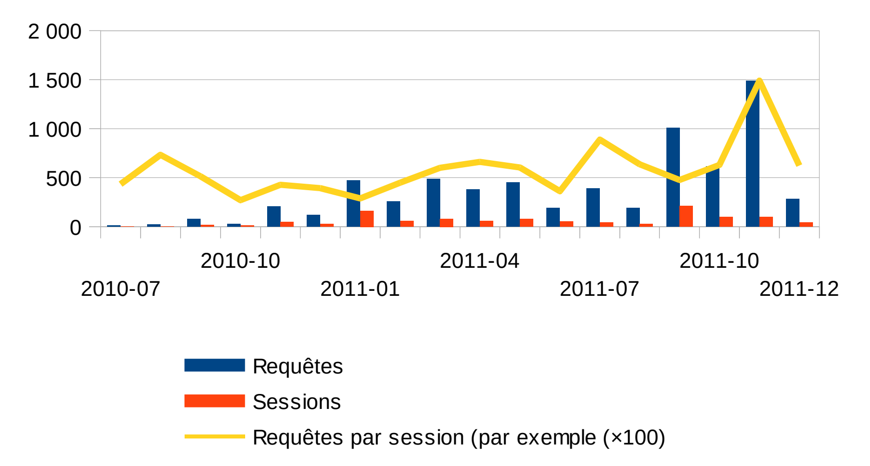


Figure 9. *Fréquentation mensuelle de l'environnement ScienQuest, pour les corpus Scientext publics, de juillet 2010 à décembre 2011*

Nous continuons à collecter des retours d'utilisateurs et à travailler sur l'ergonomie de l'interface de ScienQuest. Certaines critiques reviennent fréquemment, comme la lenteur des recherches, ou le fait que le mode de recherche simple pourrait être encore simplifié, en particulier pour les relations syntaxiques, que les utilisateurs jugent encore complexes à manipuler. Le cheminement proposé convient bien à des linguistes, et est utilisable par d'autres publics, mais présente des limites comme mentionné plus haut, en particulier pour une extension à d'autres types d'applications. Les fonctionnalités proposées (sélection de textes, concordancier, statistiques) sont relativement classiques dans la recherche en linguistique, ce qui rend notre outil utilisable (et utilisé) pour d'autres projets dans cette discipline.

Des retours récents, concernant l'utilisation de l'outil et du corpus Scientext dans l'enseignement des langues, ont montré que le guidage proposé dans notre interface n'était pas adapté à ce domaine, centré sur l'étude d'un *usage* plutôt que d'un *corpus*. Dans ce cas de figure, il faudrait réfléchir à davantage exploiter les grammaires prédéfinies du mode sémantique, et permettre à l'enseignant de préparer à l'avance des requêtes en mode libre, qui seraient mémorisées pour des parcours didactiques. Des grammaires prédéfinies pourraient également être proposées pour effectuer une annotation de certains textes particulièrement représentatifs du domaine.

Une réflexion est en cours sur l'intégration à cet environnement d'interfaces graphiques multiples, dédiées à différents scénarios d'utilisation, et en particulier en didactique des langues. Le fait que l'environnement ScienQuest s'appuie sur un service Web pour ses fonctions de recherche devrait faciliter l'implémentation du support de multiples interfaces graphiques. Il faudrait implémenter d'autres parcours plus adaptés aux autres publics, par exemple les didacticiens .

5. Conclusion

La mise à disposition de corpus annotés auprès des non-spécialistes doit passer par une réflexion sur l'ergonomie des interfaces dans la communauté du TAL si l'on souhaite vraiment que les corpus annotés sortent des placards des laboratoires. Selon nous, l'élaboration d'interfaces efficaces doit nécessairement se faire par une interaction avec les usagers non informaticiens. Dans cette perspective, nous avons proposé avec la plate-forme ScienQuest des modes de recherche simples et guidés de corpus arborés, qui rencontrent auprès des utilisateurs « grand public » un succès bien plus important que les langages de requêtes classiques. L'élaboration d'accès onomasiologiques, comme ceux qui sont proposés avec les grammaires prédéfinies, devrait permettre d'accroître le nombre d'utilisateurs des corpus comportant des annotations de haut niveau.

Outre la vitesse d'exécution, l'ergonomie du système ScienQuest doit encore être améliorée. Des évaluations fines auprès d'utilisateurs doivent être conduites prochainement. Parmi les pistes envisagées, signalons une première amélioration qui consisterait à faciliter la recherche libre et guidée en présélectionnant les relations syntaxiques en fonction des catégories grammaticales sélectionnées. Une autre piste intéressante serait de proposer un historique de la recherche facilement visualisable. Nous souhaitons également mettre en place des fonctionnalités particulièrement adaptées à une utilisation didactique, avec des scénarios spécifiques, et une interface adaptée orientée vers ce type d'application. Les nouvelles fonctionnalités seront régulièrement évaluées auprès d'utilisateurs non informaticiens. Un travail de documentation et des expériences de déploiement ont commencé, qui devraient conduire à la publication de la plate-forme sous licence *open source*. Plusieurs corpus scientifiques et généraux doivent en outre être ajoutés.

À l'heure actuelle, l'utilisation du système dépasse le cadre du projet Scientext dont il est issu. Il est par exemple utilisé en didactique du français langue étrangère (FLE) dans le cadre du projet FULS[18] et est amené à évoluer dans cette perspective. Dans le cadre du projet ANR EMOLEX[19], il intègre de nouveaux corpus en cinq langues, annotés syntaxiquement avec des analyseurs différents. Enfin, l'intégration de nouveaux corpus libres est en cours, ressources qui seront librement interrogeables en ligne : il s'agit, dans un premier temps, du corpus de l'*Est Républicain* et d'un corpus construit à partir de Wikipédia.

Remerciements

Nous remercions tout particulièrement Didier Bourigault de nous avoir permis d'utiliser le logiciel Syntex pour la réalisation de l'interface. Un grand merci aussi à Elena Melnikova qui nous a montré dans le détail le fonctionnement de Ruscorpora. Enfin, un grand merci à tous les utilisateurs de Scientext dont les retours et les évaluations nous ont été si utiles.

6. Bibliographie

Abeillé A., Clément L., Toussenel F., « Building a treebank for French », Abeillé A. (ed) *Treebanks*. Dordrecht, Kluwer, 2003, p. 165-188.

Bick Eckhard, « Live use of Corpus data and Corpus annotation tools in CALL: Some new developments in VISL », Holmboe H. (ed.), *Nordic Language Technology, Årbog for*

18. http://scientext.msh-alpes.fr/fuls/
19. http://scientext.msh-alpes.fr/emolex/

Nordisk Sprogteknologisk Forskningsprogram 2000-2004 (Yearbook 2004), Copenhague, Museum Tusculanum, 2005, p. 171-186.

Boch F., Rinck F., Nardy A., « The Evaluation of Conference Paper Proposals in Linguistics », *International Conference Writing across Boarders II,* Washington, 2011.

Boch F., Grossmann F., Rinck F., « Conformément à nos attentes… », ou l'étude des marqueurs de convergence/divergence dans l'article, *Revue Française de Linguistique Appliquée*, vol. XII-2, 2007, p. 109-122.

Bourigault D., *Un analyseur syntaxique opérationnel : SYNTEX*. Mémoire de HDR. Toulouse, 2007.

Grossmann Francis, Tutin Agnès, Garcia da Silva Pedro, « Filiation et transferts d'objets scientifiques dans les écrits de recherche », *Pratiques,* Metz, 2009, p. 187-202.

Grossmann F., Tutin A., « Evidential Markers in French Scientific Writing: the Case of the French Verb *voir* », in Smirnova E., Diewald G. (eds.), *Evidentiality in European Languages. Empirical Approaches to Language Typology* (EALT) Berlin, New York, Mouton de Gruyter, 2011, p. 279-307.

Kraif O., « Comment allier la puissance du TAL et la simplicité d'utilisation ? l'exemple du concordancier bilingue ConcQuest », *Actes des 9e Journées d'analyse statistique des données textuelles, JADT 2008*, Lyon, Presses universitaires de Lyon, 2008, p. 625-634.

Lezius W., König E., « Towards a search engine for syntactically annotated corpora ». In Schukat-Talamazzini Ernst G., Zühlke W. (ed.) : *KONVENS-2000 Sprachkommunikation* Ilmenau, Allemagne, VDE-Verlag, 2000, p. 113-116.

Mel'čuk I., *Dependency Syntax: Theory and Practice*, Albany, N.Y. , The SUNY Press, 1998.

Mertens P., « Les corpus de français parlé ELICOP : consultation et exploitation ». In Binon J., Desmet P., Elen J., Mertens P., Sercu L. (ed.) *Tableaux vivants, Opstellen over taal-en onderwijs*, aangeboden aan Mark Debrock, Symbolae, Facultatis Litterarum Lovaniensis, Series A, vol. 28. Louvain, Belgique, Leuven Universitaire Pers, 2002, p. 383-415.

Péry-Woodley M.-P., Asher N., Enjalbert P., Benamara F., Bras M., Fabre C., Ferrari S., Ho-Dac L.-M., Le Draoulec A., Mathet Y., Muller P., Prévot L., Rebeyrolle J., Tanguy L., Vergez-Couret M., Vieu L., Widlocher A., ANNODIS : une approche outillée de l'annotation de structures discursives, TALN 2009, Senlis (France), 24-26 juin 2009.

Salmon-Alt S., « Le projet ANANAS : annotation anaphorique pour l'analyse sémantique de corpus ». *Workshop sur les Chaînes de référence et résolveurs d'anaphores, TALN,* Nancy, 28 juin 2002.

Salmon-Alt S., Bick E. , Romary L., Pierrel J.M.,« La FReeBank : vers une base libre de corpus annotés », *Actes de TALN 2004*, 18-23 avril 2004.

Tutin A., Trouilleux F., Clouzot C., Gaussier E., Zaenen A., Rayot S., Antoniadis G., « Annotating a large corpus with anaphoric links ». *Proceedings of DAARC 2000* (Discourse Anaphora and Anaphor Resolution), 16-18 novembre 2000, Lancaster.

Tutin A., Grossmann F., Falaise A., Kraif O., « Autour du projet Scientext : étude des marques linguistiques du positionnement de l'auteur dans les écrits scientifiques ». *Journées Linguistique de Corpus*, 10-12 septembre 2009, Lorient.

Tutin A. (a), « Showing phraseology in context: an onomasiological access to lexico-grammatical patterns in corpora of French scientific writings ». In Granger S., Paquot M. (eds), *eLexicography in the 21st century: new applications, new challenges* . Cahiers du CENTAL. Louvain la neuve, Presses universitaires de Louvain, 2010, p. 303-312.

Tutin A. (b), « Evaluative adjectives in academic writing in the humanities and social sciences », *Interpersonality in written academic discourse: perspectives across languages and cultures*, Cambridge, Cambridge Publishing, 2010, p. 219-239.

Williams G., Millon Ch., « The General and the Specific : Collocational resonance of scientific language », *Proceedings Corpus Linguistics 2009,* University of Liverpool.